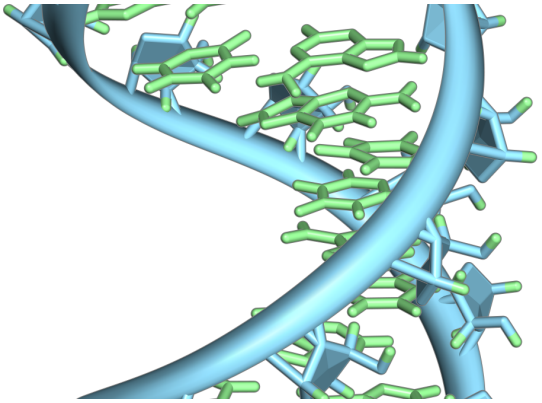
ENCODE consortium identifies RNA sequences that are involved in regulating gene expression

29 July 2020



A hairpin loop from a pre-mRNA. Highlighted are the nucleobases (green) and the ribose-phosphate backbone (blue). Note that this is a single strand of RNA that folds back upon itself. Credit: Vossman/ Wikipedia

The human genome contains about 20,000 protein-coding genes, but the coding parts of our genes account for only about 2 percent of the entire genome. For the past two decades, scientists have been trying to find out what the other 98 percent is doing.

A research consortium known as ENCODE (Encyclopedia of DNA Elements) has made significant progress toward that goal, identifying many genome locations that bind to regulatory proteins, helping to control which genes get turned on or off. In a new study that is also part of ENCODE, researchers have now identified many additional sites that code for RNA molecules that are likely to influence gene expression.

These RNA sequences do not get translated into proteins, but act in a variety of ways to control how much protein is made from protein-coding genes. The research team, which includes scientists from MIT and several other institutions, made use of RNA-binding proteins to help them locate and assign possible functions to tens of thousands of sequences of the genome.

"This is the first large-scale functional genomic analysis of RNA-binding proteins with multiple different techniques," says Christopher Burge, an MIT professor of biology. "With the technologies for studying RNA-binding proteins now approaching the level of those that have been available for studying DNA-binding proteins, we hope to bring RNA function more fully into the genomic world."

Burge is one of the senior authors of the study, along with Xiang-Dong Fu and Gene Yeo of the University of California at San Diego, Eric Lecuyer of the University of Montreal, and Brenton Graveley of UConn Health.

The lead authors of the study, which appears today in *Nature*, are Peter Freese, a recent MIT Ph.D. recipient in Computational and Systems Biology; Eric Van Nostrand, Gabriel Pratt, and Rui Xiao of UCSD; Xiaofeng Wang of the University of Montreal; and Xintao Wei of UConn Health.

RNA regulation

Much of the ENCODE project has thus far relied on detecting regulatory sequences of DNA using a technique called ChIP-seq. This technique allows researchers to identify DNA sites that are bound to DNA-binding proteins such as transcription factors, helping to determine the functions of those DNA sequences.

However, Burge points out, this technique won't detect genomic elements that must be copied into RNA before getting involved in gene regulation. Instead, the RNA team relied on a technique known

as eCLIP, which uses ultraviolet light to cross-link RNA molecules with RNA-binding proteins (RBPs) inside cells. Researchers then isolate specific RBPs using antibodies and sequence the RNAs they were bound to.

RBPs have many different functions—some are splicing factors, which help to cut out sections of protein-coding messenger RNA, while others terminate transcription, enhance protein translation, break down RNA after translation, or guide RNA to a specific location in the cell. Determining the RNA sequences that are bound to RBPs can help to reveal information about the function of those RNA molecules.

"RBP binding sites are candidate functional elements in the transcriptome," Burge says. "However, not all sites of binding have a function, so then you need to complement that with other types of assays to assess function."

The researchers performed eCLIP on about 150 RBPs and integrated those results with data from another set of experiments in which they knocked down the expression of about 260 RBPs, one at a time, in human cells. They then measured the effects of this knockdown on the RNA molecules that interact with the protein.

Using a technique developed by Burge's lab, the researchers were also able to narrow down more precisely where the RBPs bind to RNA. This technique, known as RNA Bind-N-Seq, reveals very short sequences, sometimes containing structural motifs such as bulges or hairpins, that RBPs bind to.

Overall, the researchers were able to study about 350 of the 1,500 known human RBPs, using one or more of these techniques per protein. RNA splicing factors often have different activity depending on where they bind in a transcript, for example activating splicing when they bind at one end of an intron and repressing it when they bind the other end. Combining the data from these techniques allowed the researchers to produce an "atlas" of maps describing how each RBP's activity depends on its binding location.

"Why they activate in one location and repress when they bind to another location is a longstanding puzzle," Burge says. "But having this set of maps may help researchers to figure out what protein features are associated with each pattern of activity."

Additionally, Lecuyer's group at the University of Montreal used green fluorescent protein to tag more than 300 RBPs and pinpoint their locations within cells, such as the nucleus, the cytoplasm, or the mitochondria. This location information can also help scientists to learn more about the functions of each RBP and the RNA it binds to.

Linking RNA and disease

Many research labs around the world are now using these data in an effort to uncover links between some of the RNA sequences identified and human diseases. For many diseases, researchers have identified genetic variants called single nucleotide polymorphisms (SNPs) that are more common in people with a particular disease.

"If those occur in a protein-coding region, you can predict the effects on protein structure and function, which is done all the time. But if they occur in a noncoding region, it's harder to figure out what they may be doing," Burge says. "If they hit a noncoding region that we identified as binding to an RBP, and disrupt the RBP's motif, then we could predict that the SNP may alter the splicing or stability of the gene."

Burge and his colleagues now plan to use their RNA-based techniques to generate data on additional RNA-binding proteins.

"This work provides a resource that the human genetics community can use to help identify genetic variants that function at the RNA level," he says.

More information: A large-scale binding and functional map of human RNA-binding proteins, *Nature* (2020). DOI: 10.1038/s41586-020-2077-3 , www.nature.com/articles/s41586-020-2077-3

Provided by Massachusetts Institute of Technology